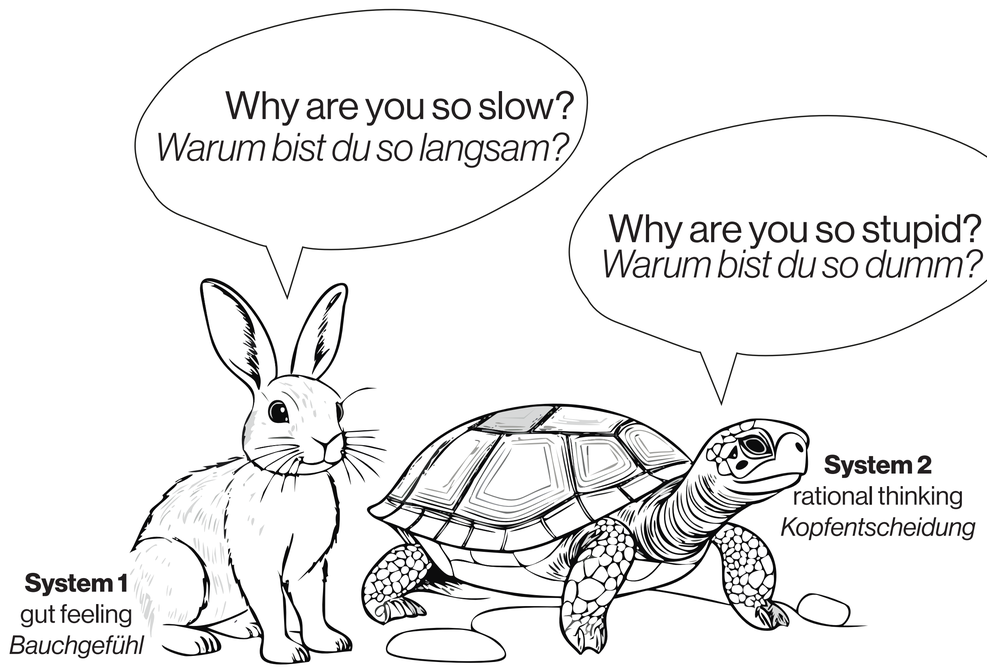
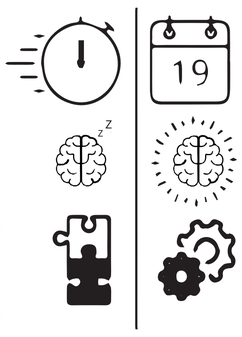
Why Large Language Models can not think?
Warum Sprachmodelle nicht denken können?



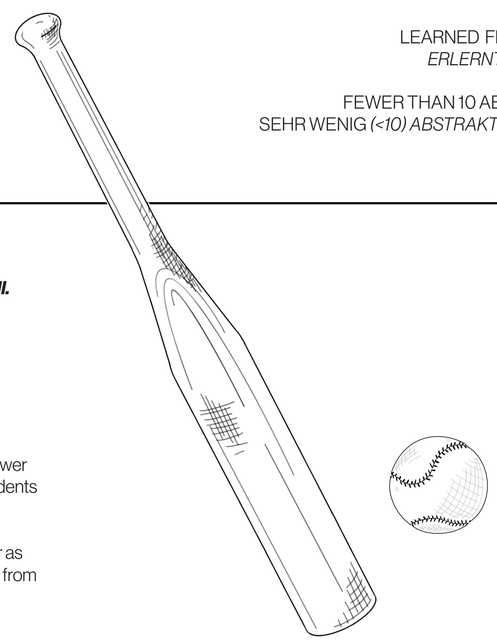
gut feeling vs. rational thinking
Bauchgefühl vs. Kopfentscheidung

The brain can be seen as a close collaboration between two systems, the fast System 1 and the slow System 2. System 1 is used for making fast, unconscious decisions based on thousands of inputs but can easily be fooled. System 1 performs routine or repetitive tasks. System 2 is used in deliberate thinking based on several single facts and can abide by the rules. System 2 makes conscious decisions and plans step-by-step processes. While it is quite effortful to use System 2, System 1 is used automatically and effortlessly.

Das Gehirn kann als enge Kollaboration zweier Systeme gesehen werden: dem schnellen System 1 und das langsamen System 2. System 1 wird für schnelle, unbewusste Entscheidungen basierend auf Tausenden von Inputs genutzt, kann aber leicht getäuscht werden. System 1 führt routinemäßige und repetitive Tätigkeiten aus. System 2 kommt bei bewussten Überlegungen basierend auf einigen wenigen Fakten zum Einsatz und kann Regeln beachten. System 2 trifft gezielt Entscheidungen und stellt schrittweise Überlegungen an. Während die Nutzung von System 2 Anstrengung erfordert, fühlt sich der Einsatz von System 1 beinahe automatisch und mühelos an.

System 1 Gut feeling *Bauchgefühl*		**System 2** Rational *Kopfentscheidung*
FAST *SCHNELL*		SLOW *LANGSAM*
UNCONSCIOUS *UNBEWUSST*		CONSCIOUS *BEWUSST*
AUTOMATIC *AUTOMATISCH*		EFFORTFUL *AUFWÄNDIG*
PATTERN RECOGNITION *MUSTERERKENNUNG*		RULE-BASED REGELBASIERT
ROUTINE OR REPETITIVE TASKS *ROUTINE ODER REPETITIVE TÄTIGKEITEN*		STEP BY STEP CONSIDERATIONS *SCHRITT FÜR SCHRITT ÜBERLEGUNGEN*
LEARNED FROM EXPERIENCE AND PRACTICE *ERLERNT DURCH ERFAHRUNG UND ÜBUNG*		CAN FOLLOW INSTRUCTIONS, LAWS & RULES *KANN ANWEISUNGEN, GESETZE & REGELN BEFOLGEN*
FEWER THAN 10 ABSTRACT FACTS ARE EVALUATED SEHR WENIG *(<10) ABSTRAKTE FAKTEN WERDEN BETRACHTET*		OVER 10,000 INPUTS AND MEMORIES ARE ANALYZED TOGETHER *<10.000 SPEZIFISCHE INPUTS UND ERINNERUNGEN WERDEN GESAMTHEITLICH BETRACHTET*

Test:



Without further ado, try to answer the following question as fast as possible.

Let us say we have a bat and a ball. Together they cost €1.10. The bat cost €1.00 more than the ball.

How much does the ball cost?

ANSWER:

Your intuitive answer probably was that the ball costs ten cents and the bat costs one euro. The correct answer is that the ball costs five cent and the bat costs one euro and five cent. Many people fail at this test, even students from high-ranking universities. So don't be discouraged, if you got it wrong, too.

This is an instance where System 1 is at work and gives an intuitive answer. Since you were asked to answer as fast as possible, your System 2 did not have time to check by the rules of arithmetic, if the proposed answer from System 1 was correct. Thus, you gave an answer which felt intuitively okay but was actually wrong.

Versuchen Sie die folgende Frage ohne lange Erklärungen so schnell wie möglich zu beantworten:

Angenommen, Sie haben einen Schläger und einen Ball. Schläger und Ball kosten zusammen €1,10. Der Schläger kostet €1,00 mehr als der Ball.

Wie viel kostet der Ball?

ANTWORT:

Ihre intuitive Antwort war wahrscheinlich, dass der Ball zehn Cent kostet und der Schläger einen Euro. Die richtige Antwort ist, dass der Ball fünf Cent und der Schläger einen Euro und fünf Cent kosten. Viele Menschen scheitern an diesem Test, selbst Studierende von hochrangigen Universitäten. Seien Sie also nicht enttäuscht, falls Sie auch falsch geantwortet haben.

Dies ist ein Beispiel dafür, wenn System 1 am Werk ist und eine intuitive Antwort gibt. Da Sie aufgefordert wurden, so schnell wie möglich zu antworten, hatte Ihr System 2 nicht genug Zeit, um nach den Regeln der Arithmetik zu überprüfen, ob die von System 1 vorgeschlagene Antwort richtig war. Daher gaben Sie eine Antwort, die sich intuitiv richtig anfühlte, aber faktisch falsch war.

Human Thinking vs. Large Language Models

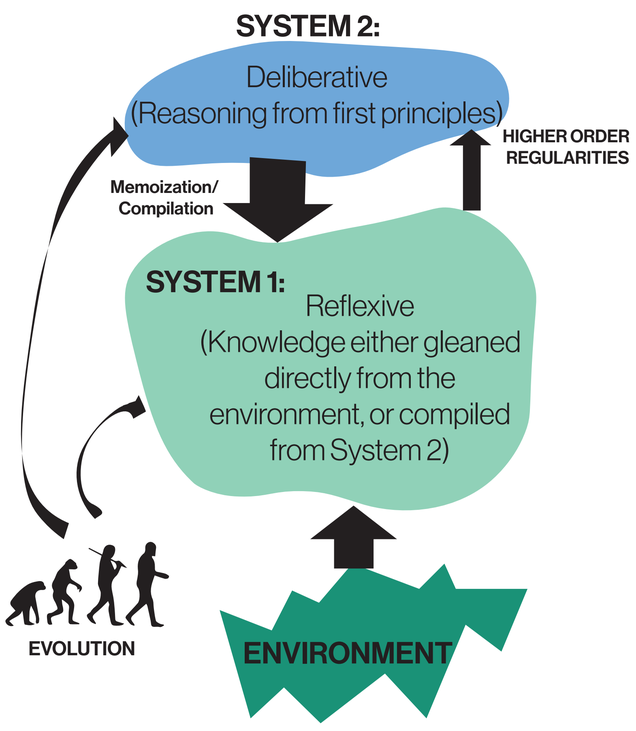
The way in which LLMs process information and generate outputs resembles System 1 thinking in humans. A LLM receives a text input and responds "immediately" by generating the response word by word "intuitively", like an unrestrained, chattering child.

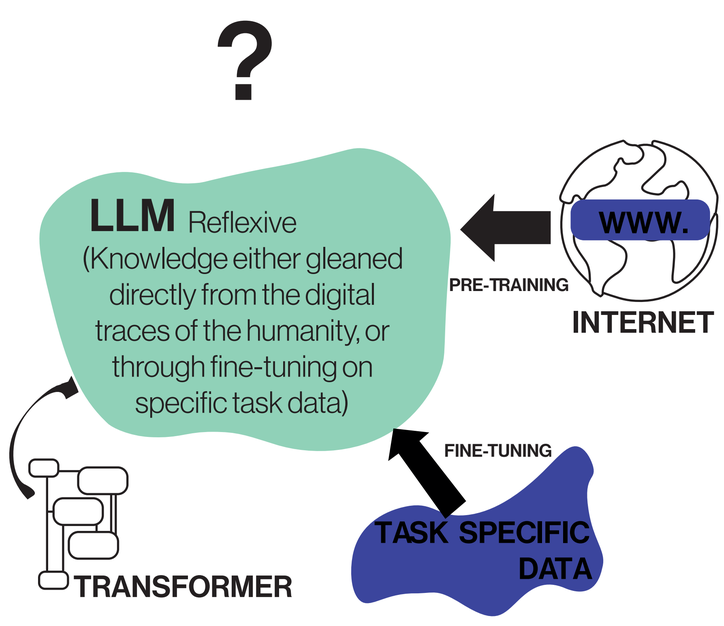
The process generates the words without a rule-based checking (System 2) that would guide the model to consider broader context such as adhering to rules or moral principles. Instead, the words are chosen based on a "feeling" of what is most likely to follow the hitherto text. One could say that there are multiple possible responses from which the LLM selects the most probable one. The crucial difference to System 2-thinking lies in the human ability to follow a rational chain of thought beyond a mere intuition. An LLM cannot do this, as an LLM does not have a System 2. Similar to the bat-and-ball problem, LLMs can therefore confidently provide incorrect answers to seemingly simple questions. The connection between classic algorithmic abilities of traditional software and the intuitive capabilities of modern LLMs is one of the hottest topics in the current AI research.

Menschliches Denken vs. Große Sprachmodelle

Die Art und Weise, wie LLMs (Large Language Models) Informationen verarbeiten und Ausgaben generieren, ähnelt dem Denken von System 1 beim Menschen. Ein Sprachmodell erhält eine Texteingabe und antwortet „sofort", indem es die Antwort Wort für Wort „intuitiv" generiert, wie ein ungehemmt plapperndes Kind.

Der Prozess generiert die Worte ohne regelbasierte Prüfung (System 2), die das Modell dazu anleiten würde, größere Zusammenhänge, wie das Einhalten von Regeln oder moralischen Prinzipien zu beachten. Stattdessen werden die Worte basierend auf einem „Gefühl" gewählt, was am wahrscheinlichsten auf den bisherigen Text folgen sollte. Man könnte sagen, dass es eine Vielzahl möglicher Antworten gibt, aus denen das LLM die wahrscheinlichste auswählt. Der entscheidende Unterschied zu System 2-Denken liegt in der menschlichen Fähigkeit über die Intuition hinaus einen rationalen Gedankengang zu verfolgen. Ein LLM kann das nicht, da das LLM kein System 2 hat. Ähnlich wie beim Schläger-und-Ball-Problem können LLMs daher voller Überzeugung falsche Antworten auf scheinbar einfache Fragen geben. Die Verbindung zwischen den klassischen algorithmischen Fähigkeiten von traditioneller Software und den intuitiven Kapazitäten moderner LLMs gehört zu den heißesten Themen in der aktuellen KI-Forschung.





MORE INFO
MEHR INFOS

Large Language Models

The algorithmic structures of **System 1 and System 2** have **evolved over time**. System 1 learns to predict the likely course of the near future through continuous experiences from the environment.

System 2 extracts rules from System 1 and supports System 1's predictions by providing abstract concepts (memoization) and modulates its intuition based on rational goals.

The **algorithmic foundation of generative LLMs** has proven to be effective after extensive research and numerous failures (a quasi-evolutionary process). The LLM learns to predict the likely course of texts from vast amounts of data from the web.

To adapt the LLM to specific tasks, such as answering questions, it is fine-tuned with **task-specific data**.

The generation of targeted abstract concepts and the rule-based monitoring of the LLM's output, akin to a **System 2, is still lacking**. This is one of the hottest current topics in AI research.

Große Sprachmodelle

Die algorithmische Struktur von **System 1 und System 2** *haben sich* **evolutionär entwickelt**. *Das System 1 lernt durch laufende Erfahrungen aus der Umwelt den wahrscheinlichen Verlauf der nahen Zukunft vorherzusagen.*

Das **System 2 extrahiert Regeln** *aus System 1 und unterstützt durch zurückgelieferte, abstrakte Konzepte (Memoization) das System 1 in seinen Vorhersagen und moduliert dessen Intuition anhand von rationalen Ziele.*

Die **algorithmische Basis des generativen LLM** *hat sich nach langer Forschung und vielen Fehlschläge (quasi-evolutionärer Prozess) als zweckmäßig erwiesen. Das LLM lernt aus umfangreichen Daten aus dem WWW den wahrscheinlichen Verlauf von Texte vorherzusagen.*

Um das LLM an gewisse Aufgaben, wie z.B. das Beantworten von Fragen, anzupassen, wird es mit **task-specific data** *zielgerichtet optimiert (=finetuning).*

Die Generierung zielgerichteter abstrakter Konzepte und die regelbasierte Überwachung des Outputs des LLMs, also ein **System 2, fehlt bisher**. *Das ist eine der heißesten aktuellen Frage in der KI-Forschung.*